Trường đại học Công nghệ thông tin

Báo cáo đồ án

XÂY DỰNG BỘ POS TAGGER CHO NGÔN NGỮ ANH

29th January 2021

Giảng viên hướng dẫn: Nguyễn Trọng Chỉnh
Môn: Xử lý ngôn ngữ tự nhiên

Võ Huy Khôi - 18520949
Lê Đoàn Nhật Minh - 18521101
Vũ Minh Luân - 18521067

Contents

1 Giới thiệu

POS tagging (gán nhãn từ loại) là một phương pháp trong xử lý ngôn ngữ tự nhiên, gán nhãn từ loại cho các từ trong một câu phù hợp với chức năng ngữ pháp của nó và ngữ cảnh trong câu.

Trong ngôn ngữ tự nhiên, một từ có thể có nhiều từ loại và điều đó gây ra khó khăn cho máy tính hiểu được ý nghĩa của ngôn ngữ. Vì vậy xác định đúng từ loại của nó là một kĩ thuật vô cùng quan trọng và là tiền đề cho các kĩ thuật xử lý ngôn ngữ tự nhiên cao cấp hơn. Ví dụ:

I fish a fish.

Ở đây từ "fish" có hai nghĩa: "câu cá" (động từ) và "con cá" (danh từ) vậy nên xác định từ loại và gán nhãn cho câu trở thành:

I/PRP fish/VBZ a/DT fish/NN

Điều này giúp cho máy tính có thể phân biệt được và hiểu được nội dung câu.

2 Ứng dụng của gán nhãn từ loại

Trong xử lý ngôn ngữ tự nhiên gồm 4 bước:

- Tiền xử lý dữ liệu
- Phân tích hình thái ngôn ngữ
- Phân tích cú pháp
- Phân tích ngữ nghĩa

Gán nhãn từ loại thuộc về bước phân tích hình thái, đây là bước phân tích các câu thành các từ hoặc cụm từ kèm theo các thông tin của từ như từ loại, phạm trù ngữ pháp và các biến thể, tiền tố hậu tố của từ.
Gán nhãn từ loại có rất nhiều ứng dụng trong việc xây dựng:

- Xây dựng các mô hình tóm tắt văn bản, phân loại văn bản.
- Xây dựng các hệ thống thông tin, hệ thống truy vấn văn bản.
- Ứng dụng trong các hệ thống dịch máy với cách thực hiện dịch dựa trên chuyển đổi.

3 Tổng quan bài toán

3.1 Mô tả hình thức

- Input: Một câu, một chuỗi các từ trong ngôn ngữ Anh và tập nhãn từ loại tiếng Anh.
- Output: Câu, chuỗi với các từ được gán nhãn.

3.2 Tiền xử lý

Phân tách kí tự thành chuỗi các từ hay còn gọi là word segmentation. Đối với tiếng Anh, các từ được phân biệt với nhau bằng một khoảng trắng, vì vậy chỉ cần phân biệt các từ dựa vào khoảng trắng.

3.3 Tìm hiểu tagset

Tập nhãn được sử dụng thường được xây dựng và phát triển từ các nhãn từ loại cơ bản trong tiếng Anh. Trong tiếng Anh, các từ loại được chia thành 2 lớp:

- Lớp từ đóng: hay còn gọi là từ chức năng (function word class) là các từ cố định, không thể mở rộng và mỗi lớp chỉ chứa một vài từ. Bao gồm: giới từ, mạo từ, liên từ, đại từ, trợ động từ.
- Lớp từ mở: Danh từ, động từ, tính từ, trạng từ.

3.4 Xây dựng kho ngữ liệu

Để tạo mô hình gán nhãn, cần phải có kho ngữ liệu để huấn luyện cho mô hình, kho ngữ liệu có thể là từ điển và các văn phạm hoặc kho văn bản đã được gán nhãn vói guidelines có sẵn.

3.5 Các phương pháp gán nhãn

- **Lexical based method:** gắn nhãn POS mỗi từ theo dạng từ xuất hiện có tần suất cao nhất trong bộ dữ liệu.
- **Rule-Based Methods:** gắn nhã POS dựa trên một quy tắc xác định. Ví dụ: trong tiếng anh, những từ có kết thúc bằng “ed” hoặc “ing” thường được gán là một động từ. Phương pháp Rule-Based Methods có thể kết hợp với phương pháp Lexical Based Methods để gắn nhãn những từ có trong bộ train nhưng không có trong bộ test.

- **Deep-learning Methods:** Sử dụng mạng nơ ron để gắn nhãn POS.
- **Probabilistic Methods:** Phương pháp dự theo xác suất. Phương pháp này gắn nhã POS dựa trên xác xuất xảy ra của một chuỗi nhãn cụ thể. Thuật toán Conditional Random Fields (CRFs) và Hidden Markov Models (HMMs) là hai phương pháp phổ biến nhất.

Trong bài báo cáo này, nhóm sử dụng phương pháp tính xác suất với mô hình Hidden Markov Models.

3.6 Mô hình Hidden Markov models

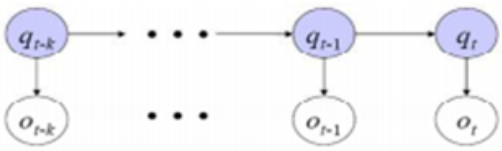

Figure 1: Mô hình minh họa chuỗi Hidden Markov

Đây là mô hình thống kê chứa các tham số quan sát được và các tham số ẩn chưa biết. Mô hình được sử dụng cho các dữ liệu có dạng chuỗi, các phần tử trước có tác động đến các phần tử ở sau.

Trong figure 1, o là các trạng thái quan sát được và q là các trạng thái ẩn tại o. Tại các trạng thái ẩn có các nút chuyển trạng thái từ q_t tới q_{t+1}.

Trong mô hình Hidden Markov ẩn bậc n, trạng thái q_t phụ thuộc vào n trạng thái ẩn trước nó, nhưng trong bài báo cáo này nhóm chỉ sử dụng mô hình Hidden Markov bigram bậc 1.

Việc dự đoán trạng thái ẩn q_t dựa vào tham số đã biết thông qua các xác suất transition (xác suất q_{t-1} xảy ra trước q_t) và xác suất emisison (xác suất o_t có nhãn q_t).

Công thức tính xác suất:

$$P(o_t|q_t) = P(q_t|o_t) * P(q_t|q_{t-1})$$

3.7 Đánh giá và so sánh

Kiểm tra độ chính xác của mô hình đã xây dựng và thực hiện so sánh với mô hình gán nhãn có sẵn khác.

4 Xây dựng mô hình gán nhãn

4.1 Bộ nhãn từ loại (tagset)

Tagset nhóm sử dụng là của bộ data Penn Treebank của NLTK. Gồm 46 nhãn trong đó có 36 nhãn từ loại và 10 nhãn dấu câu.

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	*#*
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*'s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Figure 2: Bảng danh sách tagsets

4.2 Xây dựng gold data

Gold data gồm 45 câu tiếng Anh được nhóm thu thập từ Internet. Mỗi thành viên gán nhãn 15 câu theo guideline của Penn Treebank và những từ nhập nhằng được tra cứu từ loại trên Cambridge Dictionary.

Gold data xử lý theo dạng: mỗi dòng 1 câu, các từ được phân biệt với nhãn bằng dấu "/".

Sau đó Gold data được chia thành 2 tập train, test set. Tập train gồm 36 câu, 481 từ và tập test gồm 9 câu, 122 từ.

4.3 Xây dựng mô hình Hidden Markov cho bài toán gán nhãn từ loại

Trong bộ gold data có 33 nhãn khác nhau. Gía sử có tập W = $w_1, w_2, w_3, \ldots w_n$ và tập nhãn T = $t_1, t_2, \ldots t_m$ Khi đó với mỗi từ w_i chúng ta cần tìm xác suất P = P($w_i|t_j$) (j = $0, \ldots m$) sao cho P là lớn nhất .

Vì vậy nhãn cho từ w là t_j với:

$$j = argmaxP(w|T)$$

Hay:

$$j = argmaxP(t_j|w) * P(t_j|t_{j-1})$$

P(t_j|w) tính bằng cách lấy đếm số lượng w có nhãn t_j chia tổng số lượng nhãn t_j trong bộ train.

Tạo ma trận transition TagMat có kích thước 33*33, mỗi ô TagMat(i+1,j) (hàng đầu ma trận là tượng trưng cho xác suất t_j là nhãn bắt đầu câu) tượng trưng cho xác suất P(t_i | t_j), tính bằng cách đếm số lượng nhãn t_j có nhãn trước đó là t_i chia cho tổng số lượng nhãn t_i.

Thực hiện laplace smoothing cho toàn bộ ma trận.

4.4 Sử dụng thuật toán Viterbi để tìm chuỗi trạng thái ẩn

function VITERBI(*observations* of len *T*,*state-graph* of len *N*) **returns** *best-path*, *path-prob*

create a path probability matrix *viterbi[N,T]*
for each state *s* **from** 1 **to** *N* **do** ; initialization step
  viterbi[*s*,1] $\leftarrow \pi_s * b_s(o_1)$
  backpointer[s,1] $\leftarrow 0$
for each time step *t* **from** 2 **to** *T* **do** ; recursion step
  for each state *s* **from** 1 **to** *N* **do**
    viterbi[s,t] $\leftarrow \max_{s'=1}^{N}$ *viterbi*[s',$t-1$] $* a_{s',s} * b_s(o_t)$
    backpointer[s,t] $\leftarrow \operatorname{argmax}_{s'=1}^{N}$ *viterbi*[s',$t-1$] $* a_{s',s} * b_s(o_t)$
bestpathprob $\leftarrow \max_{s=1}^{N}$ *viterbi*[s,T] ; termination step
bestpathpointer $\leftarrow \operatorname{argmax}_{s=1}^{N}$ *viterbi*[s,T] ; termination step
bestpath $\leftarrow$ the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

Figure 3: Mã giả thuật toán

Tạo ma trận có N*T chiều với T là số từ cần dự đoán, N là số nhãn khác nhau có trong bộ train ở đây N = 33.

Tại mỗi ô, xác suất được tính bằng xác suất của trạng thái trước đó nhân với max của xác suất transition của tất cả các nhãn và xác suất emission của từ. Lấy xác suất có giá trị lớn nhất và lưu lại nhãn có xác suất lớn nhất.

Khi đó có thể truy ngược lại để tìm chuỗi nhãn từ loại.

5 Đánh giá mô hình

Thực hiện dự đoán và gán nhãn cho tập test. Đánh giá mô hình bằng hệ số Accuracy: lấy số nhãn gán đúng chia cho tổng số nhãn của tập test.

Kết quả ra được là **0.5**.

So sánh với mô hình máy học khác là Conditional Random Field, đây là mô hình sử dụng thuật toán phân loại xác suất có điều kiện.

Trong thuật toán CRFs, đầu vào là tập hợp các thuộc tính (dạng số thực) từ tập dữ liệu đầu vào theo một quy tắc. Trọng số của biểu thức với các thuộc tính đầu vào cùng các nhãn đã được găn thẻ trước đó và task sẽ được dùng để dự đoán cho việc nhãn gắn hiện tại. ước lượng trọng số sao cho chỉ số likelihood của nhãn trog bộ dữ liệu train là cực đại.

Hàm features trong thuật toán sẽ xác định nhãn cho mỗi từ trong câu. Đối với tiếng Anh, có thể tìm ra các đặc điểm của mỗi từ loại để tạo thành feature, ví dụ: động từ thường có các hậu tố như ing, ed,... và danh từ là tion,...

```
'is_first_capital':int(sentence[index][0].isupper()),
'is_first_word': int(index==0),
'is_last_word':int(index==len(sentence)-1),
'is_complete_capital': int(sentence[index].upper()==sentence[index]),
'prev_word':'' if index==0 else sentence[index-1],
'next_word':'' if index==len(sentence)-1 else sentence[index+1],
'is_numeric':int(sentence[index].isdigit()),
'is_alphanumeric': int(bool((re.match('^(?=.*[0-9]$)(?=.*[a-zA-Z])',sentence[index])))),
'prefix_1':sentence[index][0],
'prefix_2': sentence[index][:2],
'prefix_3':sentence[index][:3],
'prefix_4':sentence[index][:4],
'suffix_1':sentence[index][-1],
'suffix_2':sentence[index][-2:],
'suffix_3':sentence[index][-3:],
'suffix_4':sentence[index][-4:],
'word_has_hyphen': 1 if '-' in sentence[index] else 0,
'capitals_inside': 1 if sentence[index][1:].lower() != sentence[index][1:] else 0,
```

Figure 4: Các feature được lựa chọn

Trong CRFs, cũng xây dựng dự đoán nhãn từ hiện tại theo nhãn của các từ trước đó.

Cài đặt CRF bằng thư viện sklearn_crfsuite

Accuracy đạt được là **0.713**.

Đánh giá: Mô hình cho ra kết quả không được cao và thấp hơn mô hình CRF. Nguyên nhân là các từ trong tập train chưa có độ phủ cao, có tới 481 từ nhưng chỉ có 278 từ khác nhau.

6 Kết luận

Xử lý POS tagging là một trong những kĩ thuật cơ bản và quan trọng trong xử lý ngôn ngữ tự nhiên. Hướng tiếp cận sử dụng mô hình Hidden Markov có độ chính xác chưa được cao tuy nhiên có thể cải thiện bằng cách bổ sung gold data và tăng độ phủ cho tập train.

7 Source Code

https://github.com/HuyKhoi-code/NLP/blob/main/POS_TAGGING.ipynb

8 Nguồn tham khảo

- https://trituenhantao.io/kien-thuc/nlp-xu-ly-pos-voi-thuat-toan-conditional-random-fields/
- https://tailieu.vn/doc/gan-nhan-tu-loai-tieng-viet-su-dung-mo-hinh-markov-an-2052283.html
- https://www.mygreatlearning.com/blog/pos-tagging/
- https://www.freecodecamp.org/news/a-deep-dive-into-part-of-speech-tagging-using-viterbi-algorithm-17c8de32e8bc/